SPECIAL LECTURE - Introduction to Data Processing and SAS

Beware use of the terms, "statistical difference" and "biological significance"

I. Statistics as a scientific tool

A. Integration of statistics in publications - tables, graphs, and figures

1. To show absence/presence of significant differences

2. To show, if significant, correlations and regressions

3. To show statistical modeling and prediction equations

II. Examples of where statistics may be useful

A. GROUP A: 50, 51, 49, 50, 49, 52, 50 AVG = 50 ± 0.49 (LSD = 0.00)
GROUP B: 25, 26, 24, 25, 23, 27, 25 AVG = 25 ± 0.49
STDERR and LSD indicate significant difference

B. GROUP A: 10, 80, 210, 40, 1, 4, 5 AVG = 50 ± 28.76 (LSD = 50.72)
GROUP B: 2, 48, 72, 2, 1, 26, 24 AVG = 25 ± 10.21
STDERR and LSD indicate no significant difference

C. BROWN POTATO: 0.383, 0.308, 0.394, 0.404 AVG = 0.372 ± 0.022 (LSD = 0.0293)
RED POTATO: 0.404, 0.340, 0.447, 0.415 AVG = 0.402 ± 0.023
STDERR indicates no significant difference but LSD indicates significant difference
It is acceptable to use the LSD because of "blocking" (RED values are consistently higher than BROWN values)

III. Types of statistics commonly reported in the literature

A. ANOVA (analysis of variance) - usually PROC ANOVA or PROC GLM

1. Provides "Means Squares" values and their significance

a) "Main Effects" sources of variation (i.e., TREATMENT, AGE, REP)

b) Simple Interactions (TREATMENT*AGE)

c) Nested Interactions [SPECIES(INOCULUM)]

d) Orthogonal Contrasts (PIGEONPEA vs CHICKPEA)

B. LSD (least significant difference) - part of PROC ANOVA or PROC GLM

1. Provides a number used to determine statistical differences
Other similar tests are: BON, DUNCAN, GABRIEL, SCHEFFE, SNK, TUKEY, AND WALLER

C. STDERR and STD (standard error and standard deviation) - part of PROC MEANS

1. These are typically used when very simple statistical models are used - i.e., no blocking, interactions, or nesting

D. CORR (correlation) - usually PROC CORR

1. Usually used when one or more factors in the ANOVA are significant

2. Used to determine whether or not two factors are related. For example, yield might be positively correlated with soil fertility or negatively correlated with insect damage

E. REG or NLIN (linear and non-linear regressions) - usually PROC REG or PROC NLIN

1. Usually used when factors in the ANOVA and CORR are significant

2. Used to determine the dependence of one factor on another. With this procedure, you show an equation that relates two factors and then the reliability of the equation. For instance, Lignin content (mg PLANT^{-1}) = 12.4*PAL activity (Units PLANT^{-1}) + 12.2.

IV. A simple exercise - POTATO PROTEIN EXPERIMENT OF 1995:

POTATO COLOR	REP	[PROTEIN]
Red	1	0.38763
Red	2	0.56220
Red	3	0.47174
Brown	1	0.44122
Brown	2	0.42230
Brown	3	0.50201

Write a program
that will determine
whether or not the
[protein] is different
in these potatoes

V. A more complicated exercise - PIGEON PEA / CHICKPEA EXPERIMENT

A. The first objectives will be as follows:

1. To determine whether or not certain legume*rhizobium interactions show significant differences in measurements compared to other interactions.

2. To determine whether or not wheat yield of certain legume*rhizobium interactions show significant differences compared to other interactions.

3. To determine whether or not soil nitrogen values of certain legume*rhizobium interactions show significant differences compared to other interactions.

B. There will be nine separate analyses

1. 1992 legume yields

2. 1993 legume yields

3. 1994 legume yields

4. 1993 wheat yield

5. 1994 wheat yield

6. 1995 wheat yield

7. 1992 soil samples

8. 1993 soil samples

9. 1994 soil samples

C. Generate nine data files that contain numbers for the corresponding nine programs

D. Write nine programs, similar to the example, that:

1. Input the data and perform calculations (actually, the calculations should have already been performed in LOTUS)

2. SORT and PRINT the data to make sure you are using the right data set

3. Perform an ANOVA using PROC GLM

a) The CLASS statement includes all dependent variables

b) The MODEL statement is the most important and states the relationship between dependent and independent variables. Note that it will vary from one analysis to the next

1) The model statement in the example shows the main effects (CULT REP) as well as nested [CULT(INOC), HARVEST*CULT(INOC)], and an orthogonal contrast to determine the difference between PIGEON and CHICK values. You better check the way I did the orthogonal contrast.

c) The TEST statement is needed to use the right error term with the main effects. Most main effects are tested against their interaction with REP

d) The MEANS statement is used to provide a summary of all means

e) The SORT statement is used to set up for correlations

f) The CORR statement is used to look for all possible correlations